Enabling the Discovery of Thematically Related Research Objects with Systematic Spatializations

Sara Lafia
Department of Geography, University of California, Santa Barbara, USA
slafia@ucsb.edu

Christina Last
School of Geographical Sciences, University of Bristol, UK
cl15540@bristol.ac.uk

Werner Kuhn
Department of Geography, University of California, Santa Barbara, USA
werner@ucsb.edu

Abstract

It is challenging for scholars to discover thematically related research in a multidisciplinary setting, such as that of a university library. In this work, we use spatialization techniques to convey the relatedness of research themes without requiring scholars to have specific knowledge of disciplinary search terminology. We approach this task conceptually by revisiting existing spatialization techniques and reframing them in terms of core concepts of spatial information, highlighting their different capacities. To apply our design, we spatialize masters and doctoral theses (two kinds of research objects available through a university library repository) using topic modeling to assign a relatively small number of research topics to the objects. We discuss and implement two distinct spaces for exploration: a field view of research *topics* and a network view of research *objects*. We find that each space enables distinct visual perceptions and questions about the relatedness of research themes. A field view enables questions about the distribution of research objects in the topic space, while a network view enables questions about connections between research objects or about their centrality. Our work contributes to spatialization theory a systematic choice of spaces informed by core concepts of spatial information. Its application to the design of library discovery tools offers two distinct and intuitive ways to gain insights into the thematic relatedness of research objects, regardless of the disciplinary terms used to describe them.

2012 ACM Subject Classification Information systems → Digital libraries and archives; Information systems → Search interfaces; Information systems → Document topic models

Keywords and phrases spatialization, core concepts of spatial information, information discovery

Digital Object Identifier 10.4230/LIPIcs.COSIT.2019.18

Supplement Material `https://github.com/saralafia/adrl`

Acknowledgements We gratefully acknowledge the contributions that André Bruggmann and Sara Fabrikant of University of Zurich's Geographic Information Visualization and Analysis Unit made during André's time as a Visiting Scholar at UCSB's Center for Spatial Studies as well as financial support from Jack and Laura Dangermond.

1 Introduction

In recent decades, the curation of scholarship and its access mechanisms have shifted from physical to virtual spaces. In the 1990s, physical card catalogs were migrated to online databases, trading collocation for scalability [4]. Similarly, library shelves with thematically collocated material are today largely accessed through virtual spaces, such as digital repositories organized by faceted categories [15]. This shift has increased the potential for exchange of scholarly information on the Web through semantically rich *research*

objects [5]. While online library services may provide scholars with access to millions of research objects, they do not necessarily improve the ability of scholars to serendipitously discover related objects. Such a capacity was naturally built into the physical spaces of book shelves, albeit in a limited form. Spatialization can recreate specially designed two-dimensional thematic spaces, such as neighborhoods and networks of themes. These spaces support exploration, browsing, and navigating and can be exploited in future search and discovery services, complementing standard known-item searches [10].

Exploratory search is already supported by library services, like GeoBlacklight[1] and DASH[2], which index research objects *geographically* and enable discovery and access through map interfaces. Such services curate and expose research objects based on their geographic footprints, derived from the named places that they are about (if any). They enable the integration of research perspectives by geographic locations, revealing spatial patterns, such as clusters or gaps [22]. They are especially useful in a university setting where research objects from different disciplines may refer to the same places [20]. However, geographic space only captures geographic notions of location and relatedness. Location, time, and theme are dimensions that can be used to organize observations [28] including research objects. Since in many cases, the temporal organization of research objects is comparatively straightforward (e.g. indexing research objects by their date of publication or the period they are about and displaying them using a time slider), we take on the bigger challenge of representing the relatedness of research themes.

We address this challenge by literally mapping it to the existing solution for discovery by geographic location. In other words, we ask how exploratory search for research objects can be improved by maps of thematic spaces in which related research themes are placed closer together. Conceptually and technically, we adapt our previous work to expose research objects by their geographic footprints [20] to enable discovery in specially designed two-dimensional thematic spaces, which we implement using spatialization techniques. Spatializations exploit people's familiarity with spaces in everyday life to produce intuitive visual information spaces that convey similarity through distance [18]. Spatializations, like self-organizing maps informed by cartographic principles, have been applied to efficiently visualize knowledge domains, such as the subdisciplines of geography [29]. Various types of spatializations, including point maps [24], network maps [12], and regions [11] have been proposed and empirically evaluated, demonstrating that viewers correctly interpret nearby items in abstract space as similar. Analogous to the "first law of geography" [34], this "first law of cognitive geography" states that viewers believe that closer things tend to be more similar [24].

Yet, spatialization remains underexploited, particularly in libraries, which have to deal with vast and context-dependent thematic search spaces. We see this as an opportunity to experiment with spatialization in a multidisciplinary university library repository of research objects. What further distinguishes our approach is that the spatial views we develop are designed based on core concepts of spatial information[3]; in this theory, a base concept (location), a set of content concepts (field, object, network, event), and a set of quality concepts (granularity, accuracy, provenance) capture what spatial information is about. This theory positions spatial information "at a level above data models, independent of particular application domains" [19]. We use these concepts to design two kinds of spatializations: fields of research topics and networks of research objects. A field of research topics reveals their

[1] `https://geoblacklight.org/`
[2] `https://dash.ucsb.edu/search`
[3] `https://www.researchgate.net/project/Core-Concepts-of-Spatial-Information`

distribution, while a network of research objects reveals their connectivity and centrality. We implement these two spatial views by selecting the spatialization techniques of a self-organizing map [17] and of a planar network. To obtain the necessary visual interfaces for these abstract spaces, we extend the capabilities of the same web GIS platform (ArcGIS Online) that we previously used to display and discover research objects geographically. We show how the spaces that we design are configurable and enable intuitive exploration and discovery of related research objects across disciplines.

The remainder of this paper is organized as follows. In Section 2, we present a motivating scenario to illustrate the challenge of discovering related research. Section 3 explains our conceptual approach to systematize the design of search spaces through the core concepts of spatial information. In Section 4, we implement spatializations of research objects from a university repository of masters and doctoral theses. In Section 5, we apply the spatializations, demonstrating the types of questions that they enable with examples from the previous search scenario. Finally, in Section 6, we envision discovery in spatializations informed by other core concepts of spatial information.

2 Enabling Research Discovery Across Disciplines

Discovering thematically related research in a multidisciplinary setting is both important and challenging. This is a consequence of the siloing of scientific perspectives on the world into different disciplines and the heterogeneous terminologies used within them [33]. Specifically, scholars may find it challenging to identify collaborators and methods outside of their discipline. This is problematic, given that scientific studies and applications of geographic information are increasingly transdisciplinary [19]; they may, for example, combine knowledge from sociology and psychology, or borrow methods from computer science and engineering.

As a motivating scenario, consider two published Geography theses: “Representations of an Urban Neighborhood: Residents’ Cognitive Boundaries of Koreatown, Los Angeles” [2]; and “A Temporal Approach to Defining Place Types based on User-Contributed Geosocial Content” [23]. How could the authors of these theses have gone about finding collaborators studying related topics or using related methods? Even for trained interdisciplinary researchers, disciplinary terminologies make it hard to discover related research, resulting in missed sources, insights, and opportunities for collaboration. How can researchers be made aware of thematically related research without needing to know its disciplinary terms?

A common approach to reduce mismatches in keyword-based search is to use ontologies to expand the set of search terms [3]. Such network-based approaches are often based on Linked Open Data and in the case of web journals, enable the discovery of networked data about authors, reviewers, and editors [16]. However, this approach loses the more intuitive similarity relations in the construction of terminological hierarchies [13], whose relations (e.g. broader, narrower) may not always be meaningful to the user. Thus, we propose to complement the terminological approach with an innovative spatial approach affording similarity judgments on research themes. Just as designs for successful everyday spaces, like neighborhoods and street networks, follow spatial patterns [1] and support important cognitive strategies, so can the designs for visual spaces that enable serendipitous discovery. These spatial patterns and strategies are well-understood in the geographic case (consider navigation or perspective-taking) and spatialization carries them over to abstract thematic spaces. The organizational affordances of space, well-known from geographic as well as desktop spaces, can be built into artificial spaces, creating useful and intuitive spatial structures for research themes.

3 Conceptual Approach: Making the Choices of Spaces Systematic

The core concepts of spatial information [19] offer a systematic approach to defining spatial structures by providing a typology of geographic (and other) spaces to guide the organization and interpretation of spatially referenced data. Thus, we recast spatialization as a conceptual choice of a lens through which to view data (i.e. viewing research objects as a field or network). The core concepts of spatial information provide lenses that enable distinct views on spatialized relationships, such as similarity. To go beyond purely cartographic design [22], we make our choices of spaces more systematic by basing our spatializations on those two core concepts that have a solid mathematical formalization: *fields*, formalized by continuous functions from location to theme, and *networks*, formalized by graph theory.

3.1 Choices of Spaces and their Entailments

We first review previous work to create *field* and *network* spatializations, highlighting their underlying spatial theories that inform and evolve our approach. Our thesis is that, if treated systematically and formally, there are distinct choices of spatial concepts that carry perceptual powers; these enable specific types of questions and associated insights.

Landscapes and Fields. We begin with an example from Wise's [35] pioneering intelligence work, where a spatialized display of news documents shows viewers intuitive similarity relationships based on their proximity in the display. Documents are treated as objects, with *k–means* and *complete linkage hierarchical clustering* used to project documents to a two-dimensional plane. This results in a spatialization, where the position of every news document is surrounded by a neighborhood of topics. A surface is then fit over the display, representing a terrain with peaks of high frequency terms drawn from the corpus.

While this work introduces the metaphor of a landscape or terrain to information visualization, it conflates the field of topic vectors with one of topic frequencies, essentially performing a local map algebra operation. The two field views (topic neighborhoods and topic frequencies) can be separated and an additional view of documents as objects can be added; each affords different types of reasoning (on similarity, frequency, and clustering). In our work, we will show this idea for the case of research objects. While we omit frequencies, which are not supported by adequate amounts of data, we further develop the object view into a network view that shows specific connections between documents.

Another example of an information landscape is Fabrikant's [10] spatialization of a digital library's holdings. Like Wise's approach, multidimensional scaling is used as a projection method to create a surface of keywords. However, Fabrikant's work extends the landscape metaphor by explicitly referencing three spatial concepts: 1) distance (similarity), 2) scale (level of detail), and 3) arrangement (dispersion and concentration), based on Golledge's primitives of spatial knowledge [14]. These concepts are used to systematically inform what users can do in the landscape: looking (overview), navigating (to discover items of interest), changing level of detail, selecting individual documents, and discovering relationships between documents (detail on demand). While this example moves toward conceptual formalization, it does not yet support multiple views based on different spatial concepts.

Networks and Graphs. "Maps of science" visualize research networks, ranging from co-citation networks to expertise profiles [7]. Börner et al. visualize a network of millions of university research articles embedded in an abstract spherical space. The network is rendered in a pseudo-Mercator projection, based on the idea that a Riemannian perspective, which

uses a sphere as the layout surface, offers continuous linkages. However, it is unclear what additional costs or benefits this choice imparts, as some network properties (like centrality) may be more challenging for viewers to ascertain in such a view.

The extraction of spatial and temporal information from digital text archives can inform more systematic spatializations [8]. Bruggmann and Fabrikant embed a network of toponyms (nodes) and their relationships (edges) in a geographic map to illustrate their connectivity and hierarchy. The inclusion of time in their analysis enables interesting questions about how certain places have risen or fallen in prominence over some period; this is encoded by node size (frequency of mention) along with edges (co-references with another place). The resulting networks are clear and effective, highlighting important relationships, like centrality, through systematic choices of node roles, edge roles, weighting, and embedding.

3.2 Locating Research Objects in Topic Space

Our conceptual design addresses university theses, which do not have any inherent way of locating them. We therefore model them as research objects in an n-dimensional vector space of topics. To locate them, we perform topic modeling on their titles and abstracts. Although the full text is available for most theses, we consider them to be adequately described at the metadata level; our approach gains efficiency and practicality, as only commonly available metadata are required for spatialization. Topic modeling assigns each thesis a vector of keywords (standing in for their topics) locatable in a two-dimensional topic map. We chose to assign topics to research objects, as this supports useful exploratory data analyses [6].

Field-based model. Rather than using the topic model to compute on the similarities of theses, we spatialize it into a topic map that supports visual pattern detection and similarity inferences. Our first spatialization is based on the field concept, with topics as the field attribute. Fields enable questions about the value of an attribute at any position in a given spatial and temporal domain. Field-based models underlie, but do not imply the use of, a landscape metaphor. They involve explicit choices of a spatio-temporal framework and a type of attribute (scalar, vector, spinor, or tensor).

We create a self-organizing map (SOM) using the vectors of words that result from the topic model. The SOM creates a field with a two-dimensional abstract spatial framework and a vector attribute. It represents topic locations as hexagonal cells into which point objects (representing the theses) fall. This can be seen as an example of a relative Leibniz space, generated based on objects, rather than a pre-established absolute Newtonian space [26]. The SOM satisfies the criteria for field-based models as follows:

- In its **spatio-temporal framework**, time is held constant (covering the entire period of available theses), location is controlled by the topic map, and theme is measured.
- The measured attribute **value** is an n-dimensional topic vector of words associated with the topic, ordered by their probability of occurring in theses on the topic.
- Furthermore, the topic field is **continuous**, in that a small move in position in any of six directions results in a small change in attribute value.

Network-based model. Our second choice of spatialization is based on the network concept. Networks provide views of objects that are not supported by a field view, such as questions about direct connections between objects and their centrality in the network [19]. Graphs formalize network models and give them inferential power and versatility.

Network models in general require the following explicit choices [25]: what plays the role of a node?, what plays the role of an edge?, how are edges labeled or weighted?, do they

have direction?, and is there an embedding of the nodes, edges, or both in another space? Like the field-based model, the planar network that we produce also exemplifies a relative Leibniz space. Our network spatialization of theses rests on the following choices:

- The theses (research objects) are conceptualized as **nodes**.
- The **edges** are defined based on a **binary** topical relation between theses; if two research objects have at least one of five "top topics" in common, they share an edge.
- The edges are **weighted** by the value of the topic attribute (0–1).
- The edges are **non-directed**, as topic sharing is symmetrical.
- The nodes are **embedded** in a planar space, also based on value of the topic attribute.

4 Technical Approach: Implementing Field and Network Spatializations

We spatialized masters and doctoral theses accessible through the Alexandria Digital Research Library (ADRL), a repository[4] curated by the UC Santa Barbara Library. It is named for the original Alexandria Digital Library (ADL), a project in which users could access multimedia library objects through a map interface [31]. Experimental work on ADL also resulted in a prototype "information landscape" of library objects based on frequent keywords [10]. Despite the lineage that ADRL shares with the original ADL geo-library project, it does not offer any spatial search capabilities, neither in geographic nor in thematic space; this design limitation presents an opportunity to develop spatial views that enable the discovery of research objects. We use established topic mapping and spatialization techniques [30] to:

- harvest the metadata of research theses from the ADRL repository,
- compute and assign topics to the theses using topic modeling, and
- spatialize the topics, producing a self-organizing map (SOM) and a network.

4.1 Metadata Harvesting

For our experiment, we chose research theses published by graduates of UC Santa Barbara between 2011 and 2016 that represent all 53 academic departments granting graduate degrees. The theses are accessible through a public-facing search interface, which provides keyword-based search and facets. The metadata are not accessible through an API, so we obtained permission from the UCSB Library to harvest them for the 1,731 research theses using a combination of *WGET*[5] and the Python libraries *Crummy* and *Beautiful Soup 4*[6]. The metadata follow the Portland Common Data Model[7] and are comprised of: a unique identifier; a title (of 50 words or less); a year of publication; an author; a degree grantor; a degree supervisor; a language; and a detailed abstract (no word limit) containing a problem statement, a description of methods and procedures used to gather data, and a summary of findings. Researcher contributed (uncontrolled) keywords were only available for research theses added after 2017, so we did not include keywords in our topic model.

[4] `https://alexandria.ucsb.edu/collections/f3348hkz`
[5] `https://www.gnu.org/software/wget/`
[6] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`
[7] `https://pcdm.org/2016/04/18/models`

4.2 Topic Modelling

We produced a topic model using *MALLET*[8], an open-source package developed for text-based machine learning applications. We applied Latent Dirichlet Allocation (LDA) to model the topics present in the combined text of the title and abstract of each thesis [6]. LDA is used to determine the thematic relatedness of theses by attributing the presence of each word in the combined title and abstract text to a topic characterized by a word vector. LDA is suitable, as it has been applied to many similar dimensionality reduction and classification problems [6]. LDA largely succeeds in capturing the notion of relatedness (relative to the set of inputs) despite the fact that different terms are used within those inputs (e.g., "variability" and "change" are likely to be grouped into a single topic). Thus, it is a pragmatic solution for dealing with complex notions of topics and their relatedness.

We removed the standard English stop words using a list from the MALLET package. We then experimented with between 30 to 100 topics, roughly corresponding to the number of academic departments at UCSB, which indicates a rather coarse topic granularity and targets the cross-disciplinary scope of our inquiry. We found that 71 topics provided the lowest log-likelihood value, a criterion that optimizes for the tightest possible lower bound [6]. We then assigned topic probabilities to the research objects, coded from 0 to 70. We chose to leave the topics unlabelled; they are characterized only by their word vectors. The assignment of topics provides the basis for relatedness in the following steps.

4.3 Field Spatialization

We adapted a method developed by Bruggmann to spatialize the output of a topic model [8] by using a self-organizing map (SOM) toolbox[9] for ArcGIS 9 written by Lacayo-Emery. This toolbox implements the SOM algorithm [17] in existing cartographic software, leveraging its clustering and dimensionality reduction to produce a 2-dimensional map that is readily visualized. We set the following parameters: the x / y dimension of the SOM was 42 x 42 (1,764 hexagons); the length of training was 50,000 / 500,000 runs; and the initial neighborhood radius was 42 / 6. We used the probability distribution matrix that resulted from topic modeling to produce our SOM template in ArcGIS Desktop. For cartographic readability, we only display theses from the most productive departments (those with over 50 theses). This resulted in a SOM showing 775/1,731 theses from 10 departments. Figure 1 shows the SOM, which is also published to ArcGIS Online as an interactive web application[10].

4.4 Network Spatialization

We applied a hierarchical clustering method adapted from Leicht et al. [21], which is a compromise between the single-linkage clustering method (in which a single edge is defined based on the most related pairs of nodes) and average-linkage clustering (in which an edge is defined based on the average relatedness of all pairs of nodes). We used the *tidyverse*[11] package in R to construct the edge list, assigning theses the role of nodes and shared topics the role of edges; for cartographic readability, we restrict shared topics to 5. Specifically, each thesis is characterized by the same topics and associated word vectors used to produce the SOM. For example, if *Thesis A* is characterized by Topics 2, 11, 22, 34, and 60 and *Thesis B* is characterized by Topics 4, 11, 27, 33, and 51 they share one edge based on shared Topic 11.

[8] http://mallet.cs.umass.edu/
[9] http://code.google.com/p/somanalyst
[10] http://arcg.is/0vyezH
[11] https://www.tidyverse.org/

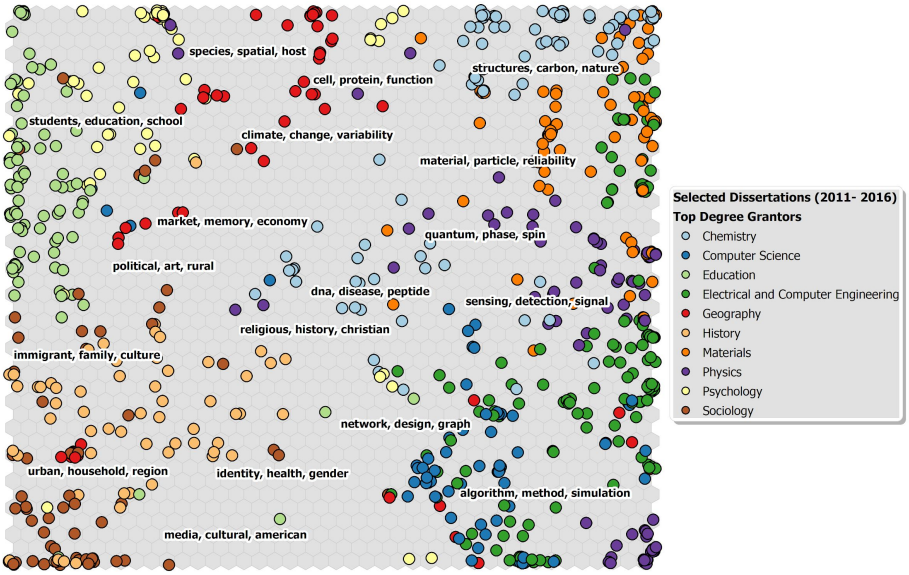

Figure 1 Theses (color coded by academic department) located in the SOM (gray tessellated topic field of themes, shown with representative terms from word vectors).

We scale node size relative to the amount that two nodes share a corresponding topic; thus, a larger node corresponds strongly with its shared topic and a smaller node does not. For example, if *Thesis A* is characterized 70% by Topic 11, its node size will be 0.7 (out of a maximum size of 1). We also embed nodes in a planar space (distinct from that of the SOM) that reflects how strongly each node corresponds to its "top-topic"; the position of each node reflects the value (0–1) of the top topic vector. To enable comparisons between the SOM and the network, we randomly sampled without replacement 775 nodes, embedded in a planar space, and connected them with edges standing in for a "top-five" topic. Figure 2 shows the network constructed with the *networkx*[12] library, which is also published in a reproducible Jupyter Notebook[13] and deployed using Binder[14].

5 Application: Discovering Thematically Related Research

The spatializations that we produce enable scholars to discover thematically related research objects, unlike the current ADRL, which does not offer any such capabilities. We apply the field and network concepts of spatial information to the motivating scenario offered in Section 2, referencing specific research objects related to the theses from the scenario. Patterns of relatedness are interpreted using Golledge's spatial primitives of *distance*, *arrangement*, and *scale* [14], which have informed previous conceptual formalizations [10].

[12] `https://networkx.github.io/`
[13] `https://github.com/saralafia/adrl/tree/master/3_network`
[14] `https://mybinder.org/v2/gh/saralafia/adrl/master`

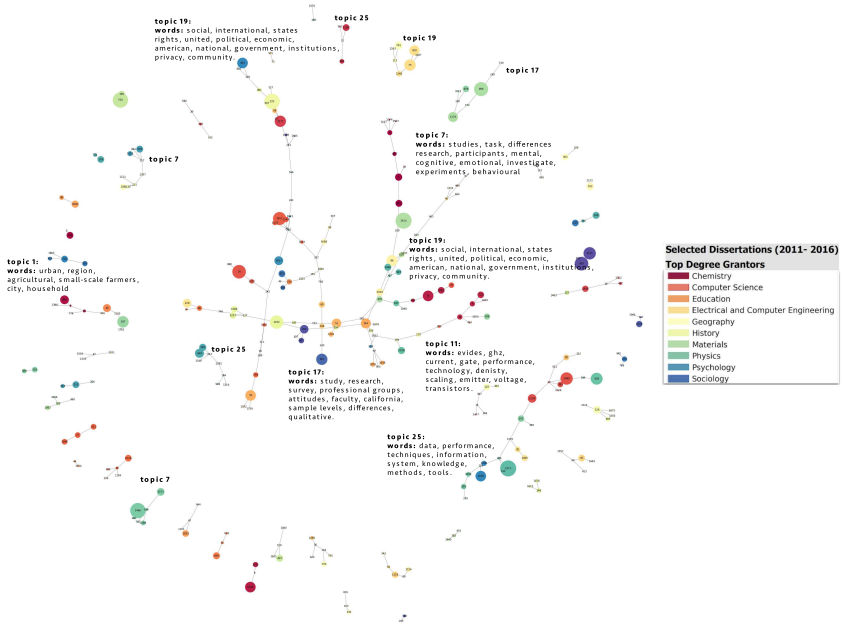

Figure 2 Theses (color coded by academic department) connected by their shared five top topics (shown with representative terms from word vectors).

5.1 Questions Enabled by a Field of Research Topics

Both the field, in the form of a self-organizing map (SOM), and the research objects used to produce it enable the discovery of related research topics. Fields enable questions about value (i.e., research topic) at a given location. A continuous field function satisfies Tobler's First Law of Geography [34], so that nearby topics in the SOM are similar. For pairs of objects, similarity can therefore be assessed by *distance*. Researchers interested in a particular area of research can see related theses by examining those closest to that area of interest in the SOM. Closely related research objects tend to fall within the area's neighborhood (i.e. a single hexagonal topic location or an aggregate of several such cells).

In the case of Bae's research from our scenario, the SOM displays six research objects from Geography, History, and Sociology within a neighborhood. Neighborhoods can be defined based on various distance thresholds. In addition to shared topics, relatedness may also reflect shared methods and techniques, as these are typically captured in abstracts as well. McKenzie's research, for example, is in a neighborhood of research objects from Computer Science and Electrical and Computer Engineering. While the subject matter of some research is different (e.g., photography or drugs), the theses share methods (e.g. "spatial, data, search…" and "learning, place, knowledge…"). Figure 3 illustrates these related research objects from the scenario, located in the SOM.

Beyond similarity of themes or methods, *arrangement*, such as the dispersion or concentration of research themes in a topic space, are also demonstrated in the field view. Theses that address the "urban, region, local…" topic are clustered and centered in the SOM, indicating that this topic pertains to many theses; conversely, topics (and their associated research objects) at the periphery of the SOM are less related to other research topics (e.g. "dna, disease, peptide…") and pertain to fewer theses. Compared with concentrated theses from other departments (like Materials, shown previously in Figure 1) the Geography department theses are dispersed; although Bae and McKenzies' theses share topics ("urban, region, local..." and "models, based, system..."), they are still distant from each other.

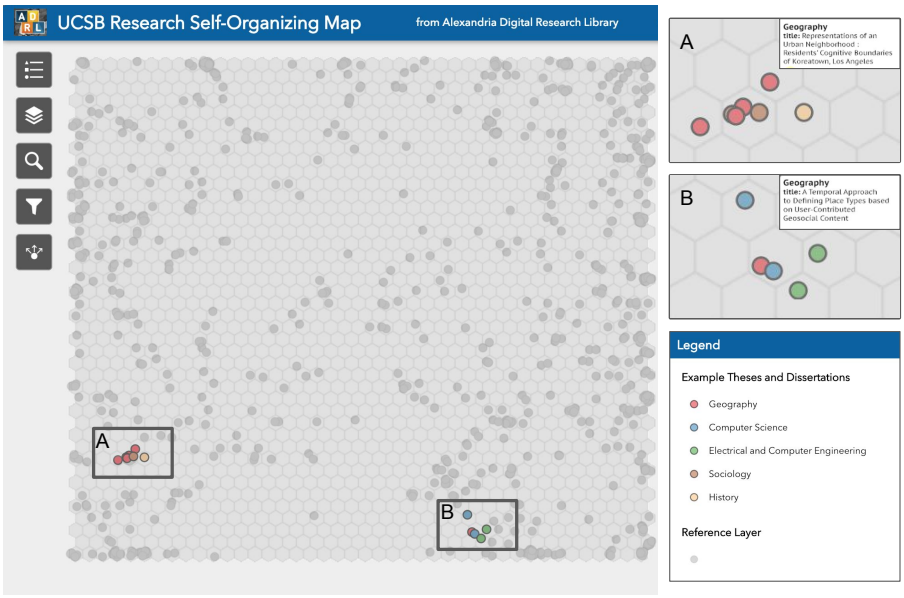

Figure 3 Selected theses (color coded by academic department) located in the SOM and surrounding: (A) Bae's geography thesis; and (B) McKenzie's geography thesis.

The field view with the thesis objects placed in it also reveals the presence and absence of research areas among existing theses. Portions of the field that do not contain any theses show research areas that may not be addressed in the repository, possibly suggesting interesting themes not yet studied and signaling opportunities for research at the boundary between disciplines. It should be noted that such gaps can also result from distortions in distance; cartogram techniques, which we have not yet applied to our field view, can be used to account for this by warping the SOM basemap [9]. Nonetheless, gaps between History and Geography surrounding Bae's research for example might suggest opportunities for integration of subject matter and techniques in this area (e.g., in the spatial humanities).

Scale in the field view is determined by topic modeling (number of input topics) and the parameters of the SOM (spatial resolution of the cells that locate topics). The size of the cells in relation to the whole field, and the dimensions of the field influence the position of topics and research objects. In our SOM, only one other thesis shares a top topic with McKenzie's research; this would likely change if the resolution of the cells changed, resulting in different topic groupings. Prevalent themes of research objects are visible at multiple levels. At the repository level shown in Figure 1, a prevalent topic appears to be about "spatial, visual, search. . . " and relates to research across many departments, including Psychology, Geography, and Computer Science. Prevalent topics of departments can also be seen from the color coding of theses by academic department (rather than by academic advisor or year of publication, which would be other possible choices).

5.2 Questions Enabled by a Network of Research Objects

Questions about the similarity, distribution, and prevalence of research topics in a repository are handled in the SOM view; however, questions about explicitly modeled relationships between the objects are not. Networks deal with these questions by encoding the relationships in their edges: for instance, are the theses of Bae and McKenzie topically related, and if so how? Figure 4 illustrates how networks convey connectivity, showing topical correspondence between departments and topical diversity within departments.

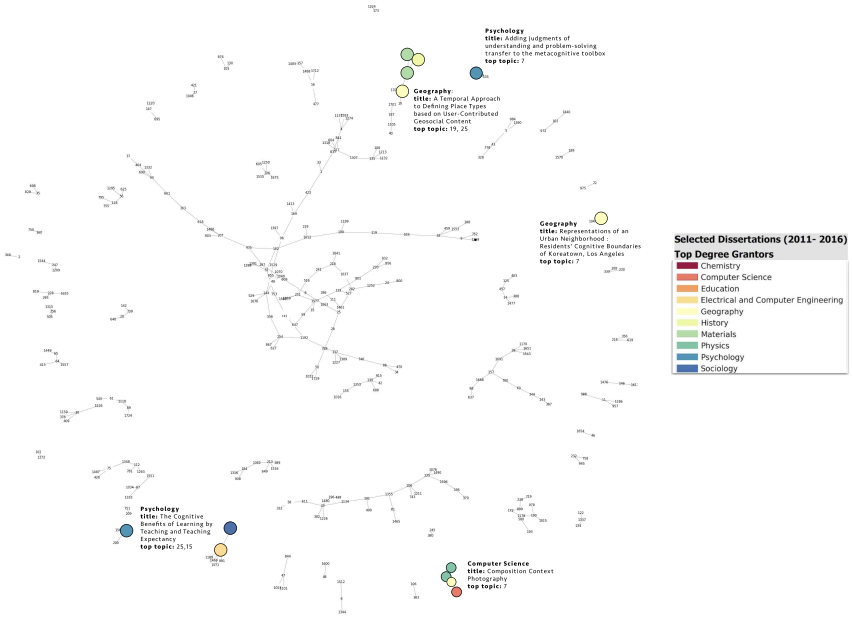

Figure 4 Selected theses (color coded by academic department and labelled by title) connected to other theses if the pair shares any five "top topics".

A network view answers questions about the specific relation encoded by network *distance*. Bae's research is most thematically related to other research objects one edge apart. A comparison between the network and the SOM shows additional similar theses, such as one from Marine Sciences, which is also characterized by the "urban, region, local..." topic.

In comparison to the SOM, where McKenzie's thesis is located next to a Computer Science thesis, there is a larger distance between them. In the network, McKenzie's thesis is closer to History and Materials theses, characterized by Topics 19 ("international, social, political...") and 25 ("data, performance, techniques..."). The Computer Science thesis shares a stronger topical relation with Geography and Physics theses, which are characterised by Topic 7 ("image, multiple, technology...").

Arrangement is related to node embedding; the most central topics in the network visualization are shared by the most research objects. More specifically, topics that intersect the central region of the network are less specific than topics that describe multiple research nodes. Niche topics are pushed toward the edge of the network; thus, theses that are heavily characterized by these topics cohere to them strongly. As shown in Figure 2, research nodes occupying a central location in the network are characterised by the generic terms "study, research, survey..." and by "studies, tasks, differences...". Conversely, theses such as those represented by the specific terms "work, material, particle..." share the fewest edges and therefore, are least central. Such theses deal with technical themes shared only by a few departments (in this case, those of Materials, Chemistry, and Electronic Engineering).

The *scale* of the network view shows a hierarchy with three levels: individual research objects, academic departments, and the repository as a whole. The nodes and the edge relations in a network can be defined in many ways. A node could represent a particular researcher and its attributes could be a list of theses published or supervised by the academic. Instead of representing a shared topic, edges could stand for a shared advisor, creating a network of "academic families or schools". While the choice to restrict edges to five top topics was pragmatic, it also illustrates the flexibility of the design approach; any kind of binary relations between research objects can be visualized.

6 Conclusion

6.1 Summary

In order to enable discovery in a multidisciplinary setting, we develop two systematic spatializations that allow users to identify thematically similar research objects. These spatializations provide a helpful alternative to known-item search by facilitating exploration; they do not require users to have prior disciplinary knowledge. To produce them, we conceptually reframe existing spatialization techniques using core concepts of spatial information. From this reframing, we produce two applications: a self-organizing map of research topics (a *field* view) and a network of connected research objects (a *network* view). In both spatializations, the relatedness of research objects can be ascertained by their *distance*; nearby topics (in neighborhoods) or objects (separated by an edge) are more related. The *arrangement* of topics and objects in each spatialization also indicates their overall relatedness; central research topics or objects tend to be more shared, while those on the periphery are niche. Finally, *scale* in both spatializations is determined during pre-processing (e.g. number of topics in the model) and spatialization (e.g. cell size; node or edge assignment). While made systematic, these choices are parameters that can be reconfigured during subsequent analysis.

6.2 Outlook

Spatializations in library services enable thematic search for research objects and complement our previous implementation of geographic search for them. Spatializing research themes extends the power of spatial search from geographically-referenced information into topic spaces, formalized in this work by core concepts of spatial information: *fields* and *networks*.

Information displays that index research by theme, location, and time [28] enable scholars to ask novel questions. The relatedness of research, indicated by proximity either in geographic location (e.g. Central American archaeology and entomology research) or thematic location (e.g. archaeological excavations of diverse ancient cultures) shows the potential for interplay between thematic and geographic views that our work enables. Furthermore, we envision allowing users to explore the spatializations in combination, gaining distinct yet complementary views of the same repository. While exploring a self-organizing map (SOM), a user can gain an overview of topics in the repository and from this, identify a specific area of interest. The subset of research objects falling into that area of the SOM can then be explored in the network, enabling further interrogation of connections. We are interested in assessing the design of our spatializations using standard usability tests, where test subjects are given questions to answer with each spatialized theme.

Temporal visualization beyond time sliders may also play a role in enabling research discovery. The meaningful representation of time-varying information [32] in disciplines like the digital humanities is notoriously fuzzy, inconsistent, and spatially variable [27]. We envision using temporal information inhering in research theses (e.g. publication date; events or periods studied) to be modeled by *events* and support reasoning on periods (time intervals). Time, made explicit and linked to spatializations, could show how research topics evolve in geographic and thematic spaces. However, events do not yet seem to provide a useful metaphor for spatializations, as they are notoriously difficult to visualize in static maps.

Visualizing the quality (as opposed to the content) core concepts of spatial information, which include *granularity*, *accuracy*, and *provenance* [19], suggests many directions for future spatialization work. *Granularity*, or level of detail, relates both to geographic scale and to the coarsened or refined topics shown in spatializations. At present, granularity

provides a clear and important intuition, as it relates directly to visualization (e.g. detail on demand). *Accuracy* relates to validity, possibly determined through comparison of multiple spatializations against domain ontologies. Finally, *provenance* may provide a way to explore the lineage of ideas (e.g. discovering related research through co-citation networks).

The long-term goals for this work are to increase awareness of relevant previous or ongoing research by applying spatial thinking to the discovery of thematically related work. Integrating research by spatialized topic, rather than siloing it by discipline, is likely to enable increased collaboration across academic disciplines. Much like browsing stacks of books in a physical library, exploring a spatialized library repository can transform a common research task into a learning opportunity or a serendipitous discovery.

References

1. Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university Press, 1977.
2. Crystal Ji-Hye Bae. *Representations of an urban neighborhood: residents' cognitive boundaries of Koreatown, Los Angeles*. University of California, Santa Barbara, 2015.
3. Ricardo Baeza-Yates and Berthier de Araújo Neto Ribeiro. *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley, 2011.
4. Nicholson Baker. Discards. *The New Yorker*, page 64–86, April 1994.
5. Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. Research Objects: Towards Exchange and Reuse of Digital Knowledge. In *The Future of the Web for Collaborative Science (FWCS 2010)*. Nature Precedings, 2010.
6. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
7. Katy Börner, Richard Klavans, Michael Patek, Angela M Zoss, Joseph R Biberstine, Robert P Light, Vincent Larivière, and Kevin W Boyack. Design and update of a classification system: The UCSD map of science. *PloS one*, 7(7):e39464, 2012.
8. André Bruggmann and Sara I Fabrikant. How to visualize the geography of Swiss history. In *Proceedings of the AGILE'2014 International Conference on Geographic Information Science,*. AGILE Digital Editions, 2014.
9. André Bruggmann, Marco M Salvini, and Sara Fabrikant. Cartograms of self-organizing maps to explore user-generated content. In *26th International Cartographic Conference*, pages 25–30, 2013.
10. Sara I Fabrikant. Spatialized browsing in large data archives. *Transactions in GIS*, 4(1):65–78, 2000.
11. Sara I Fabrikant, Daniel R Montello, and David M Mark. The distance-similarity metaphor in region-display spatializations. *IEEE Computer Graphics and Applications*, 26(4):34–44, 2006.
12. Sara I Fabrikant, Daniel R Montello, Marco Ruocco, and Richard S Middleton. The distance–similarity metaphor in network-display spatializations. *Cartography and Geographic Information Science*, 31(4):237–252, 2004.
13. Peter Gärdenfors. Semantics. In *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.
14. Reginald G Golledge. Primitives of spatial knowledge. In *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems*, pages 29–44. Springer, 1995.
15. Marti Hearst. User interfaces and visualization. *Modern information retrieval*, pages 257–323, 1999.
16. Yingjie Hu, Krzysztof Janowicz, Grant McKenzie, Kunal Sengupta, and Pascal Hitzler. A linked-data-driven and semantically-enabled journal portal for scientometrics. In *International Semantic Web Conference*, pages 114–129. Springer, 2013.
17. Teuvo Kohonen. Learning vector quantization. In *Self-Organizing Maps*, pages 175–189. Springer, 1995.

18 Werner Kuhn. Handling data spatially: Spatializating user interfaces. In *Advances in GIS research II: Proceedings of the 7th International Symposium on Spatial Data Handling*, volume 2, page 13B, 1996.

19 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012.

20 Sara Lafia, Jon Jablonski, Werner Kuhn, Savannah Cooley, and Antonio Medrano. Spatial discovery and the research library. *Transactions in GIS*, 20(3):399–412, 2016.

21 Elizabeth A Leicht, Gavin Clarkson, Kerby Shedden, and Mark Newman. Large-scale structure of time evolving citation networks. *The European Physical Journal B*, 59(1):75–83, 2007.

22 Alan M MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 2004.

23 Grant D McKenzie. *A temporal approach to defining place types based on user-contributed geosocial content*. University of California, Santa Barbara, 2015.

24 Daniel R Montello, Sara I Fabrikant, Marco Ruocco, and Richard S Middleton. Testing the first law of cognitive geography on point-display spatializations. In *International Conference on Spatial Information Theory*, pages 316–331. Springer, 2003.

25 Mark Newman. *Networks*. Oxford University Press, 2018.

26 Joan Nunes. Geographic space as a set of concrete geographical entities. In *Cognitive and linguistic aspects of geographic space*, pages 9–33. Springer, 1991.

27 Adam Rabinowitz, Ryan Shaw, Sarah Buchanan, Patrick Golden, and Eric Kansa. Making sense of the ways we make sense of the past: The PeriodO project. *Bulletin of the Institute of Classical Studies*, 59(2):42–55, 2016.

28 David Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 1978.

29 André Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5274–5278, 2004.

30 André Skupin and Sara I Fabrikant. Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2):99–119, 2003.

31 Terence R Smith and James Frew. Alexandria digital library. *Communications of the ACM*, 38(4):61–62, 1995.

32 Richard T Snodgrass. Temporal databases. In *Theories and methods of spatio-temporal reasoning in geographic space*, pages 22–64. Springer, 1992.

33 Dagobert Soergel. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, 50(12):1119–1120, 1999.

34 Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240, 1970.

35 James A Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233, 1999.